生命与健康大数据现状和展望

鲍一明[1,2*]　薛勇彪[1,2]

1 中国科学院北京基因组研究所 北京 100101
2 中国科学院大学 北京 100049

摘要　生命与健康大数据是国家人口健康和生物安全的重要基础资源。目前，我国相关数据严重流失、主权丢失、安全无法保障、再利用效率极低，亟待加快建设国家级生命与健康大数据共享平台。通过发展多元主动的数据收集新方法、互利共赢的数据共享新机制、高效智能的数据解析新技术，建立生命健康大数据的汇交、管理、共享、应用体系，维护国家数据主权，保障数据安全和加速数据应用，服务科研院所、高校、医院、企业和广大人民群众，为我国经济社会发展和人民生活改善作出重大贡献。

关键词　生命与健康，大数据，现状，展望

DOI　10.16418/j.issn.1000-3045.2018.08.014

1 人类社会面临的问题及生命与健康大数据的机遇

1.1 人类社会发展面临的诸多问题

地球已经进入到了“人类纪”（Anthropocene），人类的活动给地球的地质、生态系统造成了巨大的影响。全球变暖引起的冻土融化导致已灭绝的病原菌“重见天日”；世界人口持续增长并出现老龄化，据估计，到2020年我国65岁以上老龄人口将达1.67亿，约占全世界的1/4；全球农业生产力已经连续4年低于期望值，如不改观将不能满足地球上不断增长的人口的需求；局部冲突造成难民人数不断增加，从而引发严重的社会及经济危机；犯罪率上升、恐怖袭击、突发事件等严重威胁公共安全；重大慢病严重威胁全民健康，统计数字表明，我国有超过3.4亿的重大慢病患者，平均每分钟有8人被确诊为癌症，5人因癌症而离世。

1.2 生命与健康大数据的机遇

1.2.1 生命与健康大数据飞速增长

大数据，尤其是生命与健康大数据，将为应对上述人类社会问题起到积极的作用。生命与健康大数据是指无法在较快的时间内用传统的应用方法处理的庞大、复杂的生命与健康数据集。生命科学领域的基础研究、健康领域均产出大数据。近年来，我国生命健康方面的科技投入持

*通讯作者

资助项目：国家重点研发计划（2016YFE0206600），中国科学院信息化专项课题（XXH13505-05），中国科学院国际合作伙伴计划（153F11KYSB20160008），中国科学院率先行动“百人计划”

修改稿收到日期：2018年8月13日

续增强，国家重点研发计划启动了“精准医学研究”“重大慢性非传染性疾病防控”“生殖健康及重大出生缺陷防控研究”等重点专项，预计今后5年我国将产生300 PB以上的基因组数据。国际上，多个国家相继开展不同规模甚至百万人级的基因组测序计划。估计到2025年，全球每年将产出1 ZB的基因组数据[1]。随着健康医疗技术的不断发展，生命健康领域数据的产出越来越多。据估计，平均每个医院每年将产生665 TB的医疗数据；按此计算，仅全国1 300多家三甲医院每年就会积累约850 PB的数据。

1.2.2 健康科学的发展依赖于精准医学大数据

现代医学已经发展到基于生物信息大数据的精准医学阶段，这为恶性肿瘤、心脑血管疾病和常见病的防控和治疗提供了革命性的重大历史机遇。通过全基因组测序指导2型糖尿病治疗[2]，利用可穿戴设备收集健康大数据[3]，采用深度学习等人工智能技术帮助皮肤癌诊断[4]，运用多组学大数据整合分析进行癌症精准分型和个性化治疗[5]，以及根据DNA中包含的信息推断外貌表型、种族、地域、年龄和生活习惯[6]等，这些只是越来越多的大数据成功应用中的少数案例而已。

2 国内外生命与健康大数据的现状

2.1 国外生命与健康大数据的现状

2.1.1 国外各类基因组测序计划催生了海量的生命与健康大数据

1977年，Frederick Sanger发表的双脱氧链终止法标志着测序技术的成熟。1986年，人类基因组计划启动，并于2001年完成了人类基因组草图。2005年，454测序仪出现，下一代测序技术开始投入使用。此后，生命与健康领域的大型测序项目层出不穷，例如美国国家人类基因组研究所（NHGRI）于2003年9月启动了DNA元件百科全书计划（ENCODE Project），其主要任务是鉴定和分析人类基因组中所有功能元件。作为ENCODE项目的补充，2007年美国国立卫生研究院（NIH）启动了路线图表观基因组项目（Roadmap Epigenomics Project），该项目的目标是创建不同细胞类型的参考表观基因组图谱。几乎与此同时，欧洲的Wellcome Trust资助了千人基因组计划（1000-Genome Project）[7]。该计划由欧洲生物信息研究所（EMBL-EBI）于2008—2015年运行，主要目标是寻找在研究的人类群体中出现频率至少为1%的遗传变异。类似地，在2008年初启动的拟南芥1001基因组计划的目的是在至少1001个品系中发现相对于拟南芥参考基因组的序列变异。由美国NHGRI和NIH资助的TCGA计划[8-10]则对数千个肿瘤细胞的基因组、外显子组和转录组进行测序，试图鉴别出驱动癌症发展的公共的基因突变。NIH资助的人类微生物组计划（HMP）对生活在人类肠道和皮肤上的微生物的16S rRNA扩增子组进行测序，以期找到一组核心的、影响人类健康的微生物组。2012年，英国10万人基因组计划启动[11]。而更大的、酝酿了3年的美国政府资助的健康大数据项目100万人基因组计划已于2018年5月20日启动，该项目将建立100万人的健康大数据队列，预计耗资15亿美元，为期10年。

2.1.2 国外形成了完整的生命与健康数据中心布局

发达国家政府很早就开始重视生命与健康大数据的收集、分析和应用。早在1988年11月，美国国家医学图书馆（NLM）就意识到了“发展新的信息技术以促进对控制健康和疾病的分子过程的理解”的重要性，把Lister Hill国家生物医学交流中心的一个项目独立出来，成立了美国国家生物技术信息中心（NCBI）。从创立之初，NCBI的职责之一就是收集全世界的生物技术数据。30年来，NCBI不断发展壮大，员工数从20人增加到目前的700余人，美国国会每年拨付的经费由1990年的507.3万美元增加到2014年顶峰时的9 583.3万美元。在这个过程中，NCBI积累了全世界最大的生命与健康数据库（如GenBank、PubMed、SRA、dbGaP等）和软件资源（如BLAST、e-Utilities等），目前数据库中存储的总数据量已达30 PB，每天访问网站的用户有420万，下载数据达60 TB以上，高峰时段的点击量超过每秒7 000次。

欧洲生物信息学研究所（EBI）的前身是1980年

在德国海德堡建立的欧洲分子生物学实验室（EMBL）核酸序列数据库。1992 年，EMBL 在英国 Hinxton 建立了 EBI。EBI 最早的数据库只有两个：欧洲核酸归档库（ENA）和蛋白序列资源库（UniProt），而现在 EBI 已建成世界上最全面的分子生物学数据库集合，其管理的总数据量达 12 PB，每月用户数为 320 万人。EBI 目前有员工约 600 人，2016 年运行经费为 8 820 万美元，主要来自欧盟各国政府，特别是英国政府。

在 EMBL 和 GenBank 的邀请下，日本政府成立了日本 DNA 数据库（DDBJ）。1987 年 DDBJ 发布了 DDBJ release 1，标志着该机构开始正式运行。目前，DDBJ 的自有数据量约为 3 PB，年用户数为 268 800 人；共有约 50 名员工，年经费为 891 万美元，由日本文部省资助。

2005 年 5月，NCBI、EBI 和 DDBJ 成立了国际核酸序列数据库联盟（INSDC）。INSDC 是国际上公共领域数据共享方面最著名的组织之一，其成员每天进行数据交换，每年召开内部会议，讨论有关建立和维护序列存档的问题，并制定了一系列统一的标准和政策。INSDC 在国际生命与健康大数据收集上有着巨大的影响力，作为惯例，在主流生物医学期刊发表论文前都要将数据上传到 INSDC 成员数据库公开。

瑞士生物信息学研究所（SIB）是一个联合瑞士境内生物信息学活动的非营利性学术基金会，成立于 1998 年。SIB 的数据涵盖生命科学的不同领域，包括基因组、蛋白质组、医药健康、进化、结构生物学和系统生物学等。2017 年，SIB 核心资源被全球约 600 万用户使用，当年 SIB 管理的资金总额达到了 2 676.5 万美元。

在健康大数据领域，Epic 是美国最大的电子病历供应商，约有 1.9 亿的个人用户使用 Epic 公司的系统储存自己的电子医疗信息。Cerner 也是美国最大的电子病历供应商之一，目前，Cerner 在全世界 35 个国家支撑了 27 000 个不同大小的医疗机构。Google 的控股公司 Alphabet 旗下的 DeepMind 公司正在使用人工智能看各种医学影像，试图学会那些医生需要花上几年学习获得的经验，从而使机器学会判断病症。

2.2 国内生命与健康大数据的现状

2.2.1 国内各种类型的生命与健康大数据中心相继建成

具有代表性的包括：① 深圳国家基因库，以自产数据为主，作为节点替 EBI 收集数据。② 上海生物医学大数据中心，以中国科学院上海生命科学研究院自产数据为主，支持数据递交、发布、管理和共享。③ 微生物资源与大数据中心，以微生物资源库为主，提供微生物资源注册、查询，微生物知识查询等，用户遍布国际微生物领域。④ 国家人口与健康科学数据共享服务平台，包含约 400 个医学数据库的访问入口，以医药卫生科学数据为主。⑤ 全国公安机关 DNA 数据库[12]，于 2004 年启动，截至 2016 年 5 月 31 日，已有各类数据 4 435.8 万条，其中违法犯罪人员信息 4 071.9 万条、现场物证 149.8 万条；“打拐”DNA 数据库，累计录入人员数据 59.4 万条，DNA 数据 51.3 万条；两库数据总量达到 4 487.1 万条[12]。⑥ 北京基因组研究所生命与健康大数据中心[13-15]，数据主要来自于用户递交，数据库支持数据递交、管理、发布、共享、检索、下载、在线分析等。该数据库拥有近 100 个机构的 300 余数据递交用户，70 多个国家和地区的数据访问与下载用户，被 40 余家国际期刊认可；2018 年被生物大数据领域权威期刊 *Nucleic Acids Research*（《核酸研究》）列为与美国 NCBI、欧洲 EBI 齐名的全球核心数据中心[16]。

2.2.2 存在的问题

（1）我国缺乏生命健康大数据管理公共平台，数据流失严重。生命健康领域的期刊杂志通常要求论文的递交者把发表的数据在学界认可的数据库公开。由于我国缺乏国家层面自上而下的统一部署和规划，造成数据资源严重流失。据统计，2016 年中国大陆第一作者发表的 SCI 论文有 29.06 万篇，但其中绝大部分的数据只能被递交到 NCBI、EBI 等国际知名数据库。据估计，NCBI 数据库中 25% 以上的数据来自中国。

（2）我国缺乏生命健康大数据管理共享机制，形

成数据孤岛，利用效率低。过去的十几年里，我国通过项目经费扶持而非国家专项基金支持的形式产出了大量的数据库资源。据基于 Database Commons 数据库[①]的最新统计，我国的数据库资源总数位居世界第二；然而，大部分数据库缺少长期维护，严重缺乏深度的人工审编，数据库内容边缘化。这些因素导致大量数据库资源质量不高，利用率低，数据得不到有效共享。缺乏国家级框架的设计与部署导致我国数据库资源小而散，难以培育出处于国际领先地位的大规模优质数据中心。同样基于 Database Commons 数据库信息统计，我国引用数超过 500 次的数据库凤毛麟角，超过 1 000 次的更是为零。

（3）我国缺乏生命大数据与健康大数据的整合。生命大数据（尤其是组学大数据）与健康大数据通常是由不同主管部门下属的单位产出的。由于部门的分割及利益关系，并且缺少国家顶层的协调和制约，这两大类数据往往脱节，难以形成合力，发挥出最大效果。

3 生命与健康大数据的展望

生命与健康大数据是国家人口健康和生物安全的重要基础资源。目前，我国缺少国家级的框架与技术，对资源再利用的顶层设计、协调、管理，数据共享机制，以及长期稳定的经费支持等，这些均是制约我国生命与健康大数据研究发展的主要因素，从而造成我国数据严重流失、主权丢失、安全无法保障、再利用效率极低。因此，亟待加快建设国家级的生命与健康大数据中心，形成国家生物大数据集中管理与共享服务平台。具体来说，就是需要建成具有千万亿次计算能力和 EB 量级生物大数据储存能力的生物信息基础设施，形成能够有效承接我国生物资源、人口健康、环境与农业等大数据和支撑国家人类遗传资源有效管理的能力；建成以信息科学、生命科学、计算科学、临床医学综合交叉为基础，以云计算、人工智能等先进技术为牵引的一流生物信息平台，形成国际生物信息研究与应用开发中心。

致谢 作者感谢马英克博士对本文的编辑和整理工作。

参考文献

1 Stephens Z D, Lee S Y, Faghri F, et al. Big Data: Astronomical or Genomical? PLoS Biology, 2015, 13(7): e1002195.

2 Chen R, Mias G I, Li-Pook-Than J, et al. Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell, 2012, 148(6): 1293-1307.

3 Gao W, Emaminejad S, Nyein H Y Y, et al. Fully integrated wearable sensor arrays for multiplexed *in situ* perspiration analysis. Nature, 2016, 529(7587): 509-514.

4 Esteva A, Kuprel B, Novoa R A, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature, 2017, 542(7639): 115-118.

5 Nebbioso A, Tambaro F P, Dell'Aversana C, et al. Cancer epigenetics: Moving forward. PLoS Genet, 2018, 14(6): e1007362.

6 Vogel G. German law allows use of DNA to predict suspects' looks. Science, 2018, 360(6391): 841-842.

7 Genomes Project Consortium, Abecasis G R, Altshuler D, et al. A map of human genome variation from population-scale sequencing. Nature, 2010, 467(7319): 1061-1073.

8 Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature, 2008, 455(7216): 1061-1068.

9 Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. Nature, 2011, 474(7353): 609-615.

10 Cancer Genome Atlas Research Network, Weinstein J N, Collisson E A, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nature Genetics, 2013, 45(10): 1113-1120.

11 Turnbull C, Scott R H, Thomas E, et al. The 100 000 Genomes

① http://databasecommons.org/.

Project: Bringing whole genome sequencing to the NHS. Bmj-British Medical Journal, 2018, 361: k1687.

12 葛百川, 彭建雄, 刘冰. DNA数据库实战应用战法体系与能力建设研究. 刑事技术, 2016, 41(4): 259-264.

13 BIG Data Center Members. The BIG Data Center: From deposition to integration to translation. Nucleic Acids Research, 2017, 45(D1): D18-D24.

14 BIG Data Center Members. Database Resources of the BIG Data Center in 2018. Nucleic Acids Research, 2018, 46(D1): D14–D20.

15 Wang Y, Song F, Zhu J, et al. GSA: Genome Sequence Archive. Genomics Proteomics & Bioinformatics, 2017, 15(1): 14-18.

16 Rigden D J, Fernandez X M. The 2018 Nucleic Acids Research database issue and the online molecular biology database collection. Nucleic Acids Res, 2018, 46(D1): D1-D7.

Current Status and Prospect of Life and Health Big Data

BAO Yiming[1,2*] XUE Yongbiao[1,2]

(1 Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China;

2 University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract The life and health big data is an important resource of Chinese population health and biosafety. Currently, China's data are suffering from a severe drain and sovereignty loss, the data security cannot be guaranteed, and the efficiency of data reuse is extremely low. Thus, the construction of a national data sharing platform is urgent and should be accelerated. By developing new methods for multiple sources and proactive data collection, new mechanisms of mutual benefit and win-win data sharing and new technologies of highly efficient and intelligent data parsing, we need to establish a system for life and health big data collection, management, sharing, and application. The system will serve scientific research institutes, universities, hospitals, enterprises, and the broad masses of the people, and greatly contribute to China's economic and social development and the improvement of people's wellbeing.

Keywords life and health, big data, current status, prospect



鲍一明 中国科学院北京基因组研究所生命与健康大数据中心（BIGD）主任、研究员、博士生导师。主要从事生物数据库、病毒基因组注释和病毒进化与分类的研究。于1987年获得北京大学生物化学专业学士学位，1994年于英国John Innes中心（通过East Anglia大学）获遗传学博士学位。现为中国科学院大学健康医疗大数据国家研究院副院长，中国生物工程学会计算生物学与生物信息学专委会委员。E-mail: baoym@big.ac.cn

BAO Yiming Director and Professor of BIG Data Center (BIGD), Beijing Institute of Genomics (BIG), Chinese Academy of Sciences (CAS). His research interests include biology databases, virus genome annotation, and virus evolution and classification. He received B.S. degree from Peking University, Beijing, China in 1987, and Ph.D. from John Innes Centre (through University of East Anglia), UK, in 1994. Currently Dr. Bao is deputy director of National Institutes of Data Science in Health and Medicine, University of Chinese Academy of Sciences; and member of computational biology and bioinformatics specialized committee, Chinese Society of Biotechnology. E-mail: baoym@big.ac.cn

■责任编辑：岳凌生

*Corresponding author